Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 892–899

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Virtual Listener: A Turing-like test for behavioral believability

Arthur A. Chubarov, Daria V. Tikhomirova, Anastasia V. Shirshova, Nikolai O. Veselov, and Alexei V. Samsonovich*

National Research Nuclear University MEPhI

Abstract

Virtual Listener (VL) is a generalized prototype of a virtual character based on the principles of cognitive architecture eBICA, which uses facial expressions and "body language" (eyes movements, head rotation) to keep social and emotional contact with the user. Such contact also implies that VL needs to perceive user's facial expression and gaze, and in the long term – also intonation of the user's voice, the sentiment and content of user's speech. In this work, we present an approach to modeling a perceptive 3D virtual listener with emotional capabilities. The virtual character has a 3D face that performs real-time, realistic and believable facial expression dynamics. Our primary goal in this study was to evaluate the concept: e.g., to find out whether a virtual-agent-generated behavior can engender feelings of rapport in human speakers comparable to those that a real human listener can cause? At the same time, this article serves a limited purpose and only describes our current progress so far in addressing this question.

Keywords: virtual actor; virtual listener; emotional intelligence

1. Introduction

The long-term ambition of this research project is to analyze and synthesize human socially emotional behaviors with sufficient fidelity and control to pass a virtual-robot Turing test, to be able to create any desired virtual being with humanlike psychology, and enable total control over its virtual character type. Virtual Listener (VL) [1] is a first

* Corresponding author. *E-mail address:* avsamsonovich@mephi.ru

10.1016/j.procs.2020.02.146

step in this direction. We expect that VL should become a platform of a new type for the embodiment of human-computer interaction. In this case, the computer gets its own face and the ability to communicate through facial expressions and gaze, plus the ability to recognize the user's identity, facial expression and gaze. In this way, a new modality of human-computer interactions arises [17]. Technical tasks of the implementation of this project include both the adaptation and application of existing solutions, as well as development of new ones and attacking unsolved fundamental problems.

The first category includes emotion recognition and their expression with the facial and eyes' expressions. The second category is the problem of choosing plausible behavior, that is, the generation of adequate emotional reactions to user actions, taking into account the history of previous interaction. The second task involves the creation of a general mathematical model that describes the dynamics of a person's natural emotional reactions in various paradigms [8-10,12-15,17]. To solve it, one should accumulate a sufficient amount of data on the interaction of a person with a person in the same conditions, and then model human behavior using VL. And VL must pass a limited version of a Turing-like test, more precisely, prove to be indistinguishable from a human as a listener in the chosen virtual environment. In other words, the minimal metaparadigm of VL consists of the following. Two agents — a human participant and a Virtual Actor, or two humans — are engaged in some kind of cooperative work while looking at each other, so that each one sees the face of the other, and each can express emotions that change in real time. It is also possible to use paradigms allowing for three or more interacting agents. It is desirable that everything happens in one environment: in one virtual interior. The human participant sitting at a computer monitor or wearing a VR headset should see the partner (VL) in this interior.

This work is a continuation of the VL paper that we presented earlier [1]. We are proposing a new specific experimental paradigm. In this paper, we focus on the development of a limited Turing test in the framework of our paradigm. The implementation of the experimental paradigm involves the use of virtual and mixed reality (VR/MR).

One of the main advantages of using VR technologies can be considered a deeper immersion of the user in a virtual environment, for the simple reason of increasing the volume of the surrounding 3-dimensional virtual space. This technology allows reproducing interaction scenarios in a virtual environment with much greater credibility, which, in general, should result in a much stronger emotional response in the test subject, compared to using a monitor and traditional controls.

A virtual reality system consists of a helmet, a pair of controllers and, in most cases, special beacons to track the position of the three aforementioned user interfaces on a room scale. It is also possible to develop a virtual environment in which each participant will have their own virtual reality system using one, common tracking beacon system. Thus, the use of this system allows one to completely immerse a person in the created virtual environment and track three parts of the embedded body, namely, the head (to rotate the character's camera in virtual reality) and hands (for interactive interaction with objects inside the virtual environment). Other trackers can be added as necessary.

2. Method

2.1. New experimental paradigm

The experiment involves two participants. One participant in the experiment is wearing the VR helmet and sees in front of him two apparently identical characters-listeners, sitting on the chairs (Fig. 1). One of the characters represents the second participant in the experiment, whose facial expressions are read by a camera and reflected on the character's virtual face. The second character is VL. In this case, the participant of virtual reality wearing the helmet does not know which of them is where. All participants of the experiment continuously listen to fragments of one fascinating story (Table 1). The actual text presented to the subjects was translated into Russian and machine-converted into speech. Sentiment analysis shown in Table 1 was also done for the Russian translation, as described below.

According to the experimental paradigm, participant faces express emotions in response to what they hear. The subject wearing the helmet has the task to decide and tell by pulling a trigger, after each heard fragment of the story, which of the two character that he sees in front of him is most likely to be controlled by a human.

Fig. 1. Participant of the limited Turing test (left) and a view inside the virtual world (right).

It is assumed that the participant in the experiment makes the choice in favor of one or the other listener based on the apparent emotions that arise in her when listening to fragments of the text. We additionally register the expression of human emotions during the experiment with the help of an electromyogram (EMG) using the BrainSys Neuro-KM electroencephalograph. Therefore, the subject has to wear electrodes on her face underneath of the helmet (Fig. 2A).

Table 1. An example of a test text [11] consisting of 12 paragraphs with automatically evaluated sentiments. Values range from 0 to 100.

Text	Happy	Angry	Excited	Sad	Fear	Bored	Adequate
One morning, after chores were done, chores his boys used to do, Abraham Zimmerman told his wife through the door of their bedroom that he was going to town.	0	28	0	22	20	0	0.8
He waited a moment to see if she would ask what the job was, prepared to tell her that he was helping a crew that had gotten behind on roofing work, but she said nothing and he was relieved that he didn't have to lie.	28	0	0	23	17	0	0.7
The dogs tried to follow him down the lane.	0	36	0	18	21	0	0.8
They were confused about where Abraham's two sons had gone and so were cloying and needy.	0	0	0	32	24	15	0.7
He had to yell and throw a few rocks to get them to stay.	0	0	20	23	19	0	0.65
It was a good two-mile walk to reach the county highway along the network of gravel roads that linked the community together.	37	0	34	0	11	0	0.75
He knew that any interaction had the power to make him change his mind.	23	0	35	0	0	13	0.7
If someone asked him to lend a hand with something, he wouldn't be able to refuse. But he met no one.	0	39	0	15	0	13	0.7
He hadn't been up Cording Road since the evening of the accident, but because the accident had everything to do with his decision this day, it seemed necessary to pass the spot where his boys died.	22	0	0	43	16	0	0.6
There was no visible sign, and he resisted wading into the ditch grass to search for one.	0	27	0	27	19	0	0.55
And then he saw, on the fence, the remnants of a bouquet someone had tied there with twine.	47	0	0	16	15	0	0.5
Anyone who didn't know about the accident would assume it was just a tangle of wildflowers blown off a windrow after haying.	0	33	0	0	22	17	0.6

The technology of electromyography [6, 7] becomes necessary in studies of human emotional state dynamics [17], especially when it is not possible to use camera-based methods of reading facial expressions. The idea is to transmit messages via the Sockets protocol in real time from the BrainSys program that acquires the EMG data to another

program, which performs data analysis and sends the extracted information about expressed emotion to the VR simulator (Unity) for appropriate subsequent processing. This is done again using the Sockets interface. Thus, it is possible to transfer the necessary values to the local entities in VR, as well as to replicate these values on all devices participating in the session.

A

B

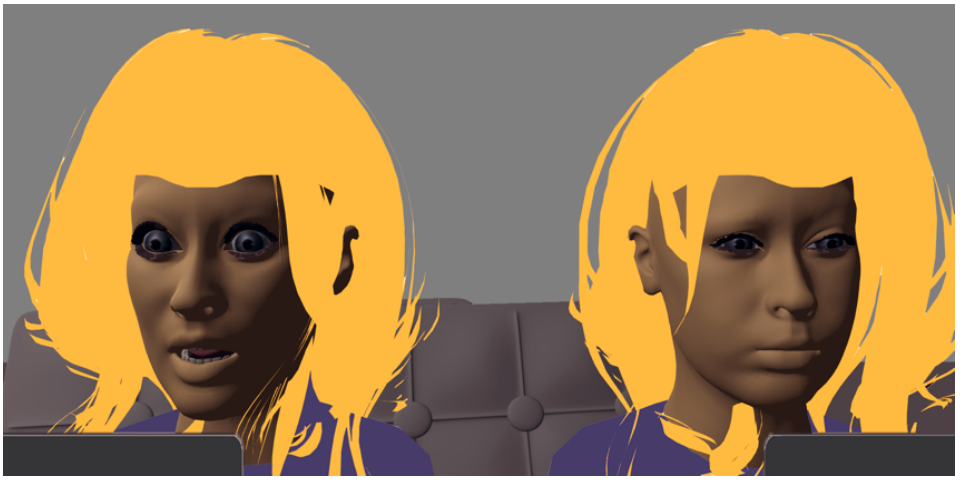

Fig. 2. A: The participant is wearing facial electrodes for electromyography below the VR helmet. B: Two virtual characters-listeners exhibit different emotional states in response to the same text fragment.

The experiment on passing the limited Turing test in the framework of the new paradigm involved 16 volunteers (n = 16, 13 male and 3 female, all age 18 to 22). All of them were students of the National Research Nuclear University MEPhI.

2.2. From text to facial expressions

For the sentiment analysis of text, we used ParallelDots Emotion analysis [2]. ParallelDots is a machine learning platform which provides powerful AI-driven solutions with REST API [3] for natural language processing problems. It supports 14 languages including English, German, French etc. The most important for us was the Russian language support, because we use Russian text in our experiment.

ParallelDots service uses Convolutional Neural Networks (CNNs) [4] to detect emotions in texts. CNN consists of convolutional and pooling(subsampling) layers followed by non-linear activation function. A convolution is the application of learnable filter(kernel) to an input. Result of repeated filter application called a feature map - representation of features detected in input data. Pooling layers subsample their inputs. It helps to reduce feature map size, but keeps most of the useful information. It is like a detecting specific feature in input data. Since CNNs can only work with matrices, input text have to be represented as a numeric matrix. Each row of the matrix corresponds to one token - a word or a character. That is, each word represented as a vector, typically as word embeddings [5]. Despite the fact that the CNN lose information about the order of words, it has large receptive field and could “look” on the big part of sentences at once. This makes CNN a good solution for classification problems, including emotion detection.

ParallelDots uses API keys to authenticate requests to text analytics APIs. In order to get one we registered on their website and created account. Emotion detection API accepts POST request containing API key and text for analysis. It returns JSON response containing scores for six basic emotions (Happy, Angry, Excited, Sad, Fear, Bored) (Fig. 3).

The score is then normalized and interpreted as probabilities that the listener will express a particular emotion in response to the text. Scores were computed for every sentence and paragraph of our text (Table 1). Then for each sentence three most likely emotion classes were selected.

Facial expressions of VL were formed based on these scores. To validate the created facial expressions, we asked participants to separately assess the expressions after the experiment, using standard affective scales.

Using 3D graphics, we created realistic 3D characters and displayed computed emotional states on virtual 3D faces.

Thus, in this work we use one of the facial recognition techniques that offers the highest recognition accuracy and sufficient coverage of possible configurations of human facial expressions. Such a tool can be considered the face recognition module of the ARKit platform. This platform uses the True Depth Camera of iOS devices, which allows three-dimensional scanning for the production of applications with the capabilities of augmented reality. In general, augmented reality offers a way to add two-dimensional or three-dimensional elements to the camera image in real time, so that the added content complements the real-world image.

ARKit simultaneously uses the capabilities of tracking the position of the device and special algorithms for recognizing individual elements from the data coming from the camera to simplify the creation of applications using these capabilities. ARKit uses the True Depth Camera to provide real-time developers with information about the configuration of the user’s face, which can be used to locate their own content. Thus, one can use this data to create a realistic mask according to these parameters that repeats the user’s mimicry.

ARKit provides the ability to manipulate the Virtual Actor model through using a set of special parameters called BlendShapes. These parameters are used to control the animation of two-dimensional or three-dimensional models in real time. There are 52 different BlendShapes coefficients, while the developer can use any subset of them at any time.

Thus, using BlendShapes, it is possible to synchronize the models of actors of two subjects in VR by synchronizing them with the help of the multi-user component of the BlendShapes set, as well as to produce their own synthesis of facial expressions on in real time. This type of data is a simple dictionary, each key of which represents one of the many face parameters recognized by the platform. The position corresponding to each key is described by a floating-point number from 0.0 to 1.0, where 0.0 corresponds to the minimal intensity and 1.0 to the maximal intensity. Since ARKit does not introduce its own restrictions on the use of these parameters, one can use any subset of them or all at once in real time. It is also possible to record face configurations and subsequently to display them at any time.

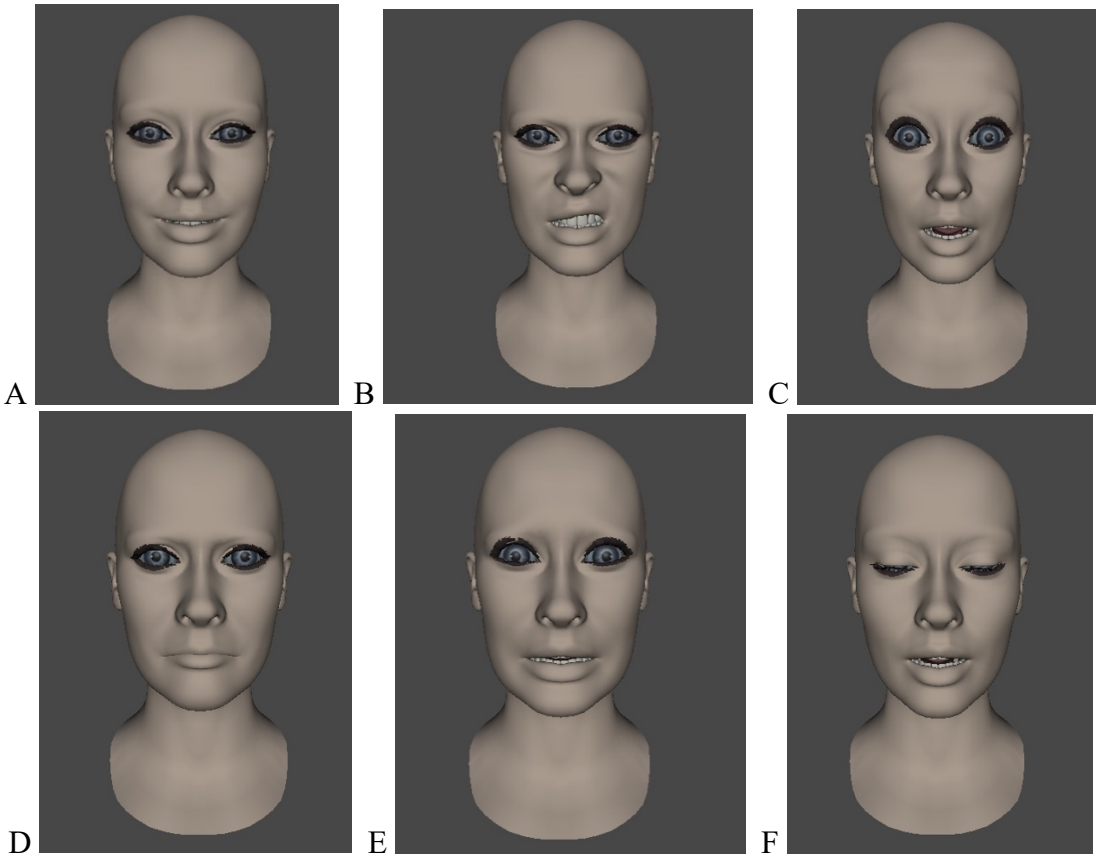

Fig. 3. Various configurations of blendshapes of the final model representing specific emotional states. A happy, B disgusted, C surprised, D sad, E angry, F bored. More accurate emotional values of these facial expressions were obtained as vectors on a semantic map by polling the subjects.

3. Preliminary Results and Analysis

The result of the experiment was a nearly exact match (52% to 48%) of the total numbers of wrong and correct trigger pressings (out of the 166 total events of trigger pressing), meaning that the limited Turing test was successfully passed: no statistically significant differences were found over all participants in the experiment between perception of the left and the right characters ($p>0.68$). All data are summarized in Table 2.

Table 2. Results of the limited Turing test. Swap means that the human and VL switched avatars.

Participant ID	Swap	Left trigger	Right trigger	Correct?
1	0	6	8	0
2	0	4	9	0
3	0	5	7	0
4	0	4	10	0
5	0	4	10	0
6	0	8	6	1
7	1	11	3	0
8	1	5	10	1
9	0	8	6	1
10	0	5	9	0
11	1	3	11	1
12	1	3	11	1

In other words, the virtual character controlled by the model was not perceived differently from the character controlled by a human participant. The result was the same when judged by the final answers of all participants, determined by the majority of right or wrong trigger pressings for each participant. The score in this case was 7:5 (Table 2, the right column), with 7 wrong and 5 correct overall participant opinions. Again, the difference is not significant (Exact Fisher Test $p > 0.68$).

4. Discussion

According to these results, we can tentatively talk about passing the limited Turing test. Indeed, the participants in the experiment could not unambiguously give preference to one of the listeners, in the role of one of which was a person (another subject), and in the role of the other - the program. However, we cannot reliably assert that the choice of the subjects was determined by the match of the emotional flavor (sentiment) of the text fragments and the emotions that the characters' faces demonstrated. The observed outcome could be due to the limited expressibility of the selected method of facial expression transmission. In fact, the data recorded from the human face were passed through a narrow "bottleneck" of binning in time and in the affective space, and therefore could not be perceived as natural human behavior on the other end.

On the other hand, the obtained result may not be surprising at all in the context of the related work [12,17]. Another potential problem is the mirror vs. two selves dilemma in perception of the two similar characters [16].

In the future, it is planned to conduct a correlation study, the purpose of which will be to confirm the relationship between the sentiments of the fragments of the text, the facial expression of the subject's emotions and the emotions that the selected character demonstrates. This approach will allow us to improve the model and to construct an algorithm that will yield adequate emotionally intelligent behavior of a Virtual Actor in response to spoken text.

This work continues the VL project that we presented earlier [1] and is still in progress. VL is a special kind of a virtual actor - an intelligent assistant and a partner, the main role of whom is to establish and maintain a socially emotional contact with the participant, thereby providing a feedback to the human performance, using minimal resources, such as facial expressions. This sort of a personal assistant is intended for a broad spectrum of application paradigms, from assistance in preparation of lectures to creation of art and design, insight problem solving, and more, and can be expanded to assist human users in many professional tasks and procedures.

Acknowledgements

This work was supported by the Russian Science Foundation Grant # 18-11-00336. The authors are grateful to all participating MEPhI students and to the MEPhI Academic Excellence Project for providing computing resources and facilities to perform experimental data processing.

References

[1] Alexander A. Eidlin, Arthur A. Chubarov, Alexei V. Samsonovich: Virtual Listener: Emotionally-Intelligent Assistant Based on a Cognitive Architecture.

[2] ParallelDots: https://www.paralleldots.com/emotion-analysis. Last visited: October 12nd, 2019.

[3] Fielding, R. T., & Taylor, R. N. (2000). Architectural styles and the design of network-based software architectures (Vol. 7). Doctoral dissertation: University of California, Irvine.

[4] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[6] Fridlund A.J., Cacioppo J.T. Guidelines for human electromyographic research. // Psychophysiology. 1986. Vol. 23. No. 5. P. 567-589.

[7] Golland Y., Hakim A., Aloni T., Schaefer S., Binnun N. L. Affect dynamics of facial EMG during emotional experiences continuous // Biological Psychology. 2018. P. 7-22. doi: 10.1016/j. biopsycho.2018.10.003.

[8] Samsonovich A. V. Emotional biologically inspired cognitive architecture // Biologically Inspired Cognitive Architectures, 2013. Vol. 6, p. 109-125.

[9] Samsonovich, A.V. (2018). Schema formalism for the common model of cognition. Biologically Inspired Cognitive Architectures, 26: 1-19.

[10] Samsonovich A. V. On the semantic map as a key component in socially-emotional BICA // Biologically Inspired Cognitive Architectures, 2018. Vol. 23, p. 1-6.

[11] Smith, A. (2019). Late Rumspringa. Narrative (online resource). Retrieved in September 2019 from https://www.narrativemagazine.com/issues/fall-2019/fiction/late-rumspringa-austin-smith

[12] Yalcin, O.N. and DiPaola, S. (2020). M-Path: A Conversational System for the Empathic Virtual Agent. *Advances in Intelligent Systems and Computing*, vol. 948, pp. 597-607. Cham, Switzerland: Springer.

[13] Samsonovich, A.V. (2020). Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. Cognitive Systems Research, 60: 57-76. DOI: 10.1016/j.cogsys.2019.12.002

[14] Samsonovich, A.V. & Ascoli, G.A. (2013). Augmenting weak semantic cognitive maps with an "abstractness" dimension. Computational Intelligence and Neuroscience, Volume 2013, Article ID 308176. DOI: 10.1155/2013/308176

[15] Samsonovich, A.V., De Jong, K.A., Kitsantas, A., Peters, E.E., Dabbagh, N., & Kalbfleisch, M.L. (2008). Cognitive constructor: An intelligent tutoring system based on a biologically inspired cognitive architecture (BICA). Frontiers in Artificial Intelligence and Applications. Vol. 171, Issue 1, pp. 311-325.

[16] Samsonovich, A.V. & Ascoli, G.A. (2005). The conscious self: Ontology, epistemology and the mirror quest. Cortex, 14 (5): 621-636. doi:10.1016/s0010-9452(08)70280-6

[17] Tikhomirova, D.V. Chubarov, A.A., & Samsonovich, A.V. (2020). Empirical and modeling study of emotional state dynamics in social videogame paradigms. Cognitive Systems Research, 60: 44-56. DOI: 10.1016/j.cogsys.2019.12.001